MaHPIC File Naming Standards
February 23, 2018

Rationale

Over its five years, it is expected that the MaHPIC project will generate about a quarter million files. Effective file naming standards are critical for the organization of the Data Repository and the usability of custom data collections for Modelers and other MaHPIC researchers.

A clearly defined directory structure cannot be maintained once a collection of ad hoc files is downloaded. For example, in year 4, an investigator may wish to download all RNA-seq expression files that were generated for *M. mulatta* (5 subjects), not infected (control), and infected (separately) with 3 different pathogens. The user may conceivably download > 100 files that tabulate read counts per gene. There can be no clearly defined directory structure for this collection of files that would describe the origins of each file. The user will have to rely on the names of the files to proceed with analysis.

Driven by these user needs, a need for both human readability and programmatic parsing, and balancing with file name length, the Informatics Core will adopt a consistent file naming strategy for all files in the Data Repository. ‘X’s will be used to fill in fields where information is missing or limited.

MaHPIC file names

Every file name will contain three defined sections, separated by underscores. Two or Three letter codes have been developed for use in the file names, for Data producers, Species, Specimen types, and Countries.

- The first section will contain clearly defined fields, with fixed lengths, that will accept a pre-determined range of values and codes. This section contains information about the experiment, the host and pathogen, and the specimen types.
- The second section will contain file specific information such as a date, or algorithm used, or type of data contained within, etc
- The third section will contain the LIMS barcode that corresponds to the sample from which this file was produced.

<Exp#><TP><DataProducer><HostPat><SpecimenType>_<FileSpecificInfo>_<LIMSBC>.ext

(<> define fields, and will not be part of file names)

#	Field	Description	Length and type expected	Example
1	Exp#	Experiment ID, Complimentary Exp ID, Supporting Exp ID,	3 to 4 – alphanumeric Possible 4th character if experiment is divided into parts ‘A’ and ‘B’	E04 or C04 (for complimentary projects) or S04 (for supporting experiments) E07A and E07B use a 4th character
2	TP	Time point	3 - alphanumeric	T06 or D13 (for dailies)
3	DataProducer	Data producer code	2 - alphabetic	FG (for Functional Genomics)
4	HostPat	1 host and up to 2	6 - alphabetic (2 for	MmCyXX (for M. mulatta

		pathogens or drugs codes	each organism)	infected with P. cynomolgi)
5	SpecimenType	Sample / Specimen type code	2 - alphabetic	IR (for infected red blood cells)
6	FileSpecificInfo	Information that is specific to this particular file	Variable	RawQ20LocalAlignBWA
7	LIMSBC	LIMS barcode	7 - numeric	1234567
8	ext	File extension	Variable - alphanumeric	bam / xlsx / wiff

Example filename:
E04T01FGMmCyXXIR_RawQ20LocalAlignBWA_1234567.bam
(Colors above delineate separate fields)

Special characters in the file names

Since the fields, their order and lengths are pre-determined in the file naming standard, to enable programmatic parsing, **if no value can be assigned** to a field in a file name, the characte**r X** must be used as a **place holder**.
For example, if a specimen is extracted from a host that was infected with just one pathogen, XX must be used to fill the <HostPat> field => MmCyXX (*M. mulatta* infected with *P. cynomolgi*)

On the other hand, **if multiple values can be assigned** to a field in a file name, special codes must be used to represent this case.
For example, for a file that aggregates the results from several files (each one of them resulting from an individual barcode), the code **MULTIPL** must be used in the <LIMSBarcode> field.

X	Filler / place holder - when no value can be assigned to a fixed field. Character X (upper case) is to be repeated as many times as the length of the fixed field
E99	when data from **multiple Experiments** are contained in a file
M99	when data from **multiple Time Points** were used to generate this file
P##	when daily samples are collected before inoculation and '##' = pre-inoculation collection day
D##	when daily samples are collected after inoculation and '##' = post-inoculation collection day
Nec	when sample is from Necropsy and is neither daily or time point note that some files contain 'N00', an older designation that cannot be changed
ZZ	when data from **multiple specimen types** were used to generate this file
MULTIPL	when data from **multiple LIMS barcodes** were used to generate this file, for example, FlowJo files, Excel files.
-	Hyphens can be used to separate information within the file specific info field

Data Producer Codes

Data Producer	**2 Letter Code**
Functional Genomics Core	FG
Immune Profiling Team - Adaptive	AI
Immune Profiling Team - Innate	II
Immune Profiling Team (if ‘Adaptive’ or ‘Innate’ are not appropriate)	IM
Informatics Core	IN
Lipidomics Core	LI
Malaria Team - CDC	MC
Malaria Team - Emory	ME
Malaria Team International	MI
Metabolomics Team	MB
Modeling Team - Gutierrez	MG
Modeling Team - Styczynski	MS
Modeling Team - Voit	MV
Pathology Team	PT
Proteomics Team (CDC and CCRC)	PR
Proteomics Team (SRI)	PS
Quantitative Metabolomics Team (Biocrates)	QM
Targeted Proteomics Team (SOMAscan)	TP
Telemetry Team	TE
Yerkes Sequencing Core	YS

Organism / Species / Drug Codes

Species	**2 Letter Code**
Aotus nancymaae	An
Pyrimethamine	Dp
Homo sapiens	Hs

Macaca fascicularis	Mf
Macaca mulatta	Mm
Plasmodium coatneyi	Co
Plasmodium cynomolgi	Cy
Plasmodium falciparum	Fa
Plasmodium knowlesi	Kn
Plasmodium vivax – P01	Vp
Plasmodium vivax – Sal1	Vi
Saimiri boliviensis	Sb
Drug	**2 Letter Code**
Artemether	Da
Biotin	Bl
Coartem	Dc
Chloroquine	Cq
Primaquine	Dq
Quinine	Qn

Sample / Specimen Codes

Specimen	**2 Letter Code**
Axillary Lymph Node	AL
Blood Clot	BT
Blood Pellet	BP
Blood Platelet	PT
Bone Marrow	BM
Bone Marrow Cells	MC
Bone Marrow Lymph	BL
Cell Pellet	CP
Cerebrospinal Fluid	CF
Cryopreserved Plasma From LymphoPrep	CR

Dander	DA
FACS (Fluorescence-Activated Cell Sorting)-Sorted	FC
Genomic DNA	DN
Infected Red Blood Cells	IR
Inguinal Lymph Node	IL
Metadata	MD
Packed Red Blood Cells	PA
Peripheral Blood Mononuclear Cells	PB
Plasma	PS
Platelet	PL
Platelet From Capillary Samples	PC
Platelet Rich Plasma	PR
Rectal DNA	RD
Rectal Swab	RS
Red Blood Cell	RC
Red Blood Cell Membrane	MN
Saliva	SA
Serum	SE
Spleen	SP
Splenic Lymph Node	SL
Stool	ST
Thyroid Gland	TH
Tissue	TI
Urine	UX
Whole Blood	WB
Whole Blood Capillary	WC
Whole Blood PCR Pellet	WP

Human Host

Collaborators from several countries will provide samples, from humans with malaria infections, as well as controls.

In the place of standard experiment codes 'E##', there are specific codes for human experiments that begin with 'Hu' and end with a letter indicating the country and year of sample collection. These are described in a special table below. These were created to simplify the naming of files from human samples because each human dataset includes one or more standard MaHPIC Experiment IDs (E##):

- E08 = samples from humans infected with *P. falciparum*
- E09 = samples from humans infected with *P. vivax*
- E10 = samples from humans infected with *P. vivax* and *P. falciparum*
- E11 = samples from humans infected with *P. knowlesi*
- E12 = samples from uninfected humans

Usually, there will be no concept of time points for human samples. On the other hand, keeping track of the country of origin is important. So, the <TimePoint> field will be replaced with a three-letter country code, for Human samples. If there is information about cities that needs to be included in the file names, the <FileSpecificInfo> may be used.

<HumanExpID><CountryCode><DataProducer><Host-Pathogens><SpecimenType>_<FileSpecificInfo>_<LIMSBarcode>.<ext>

#	Field	Description	Length (type expected)	Example
1	HumanExpID	Human Experiment ID indicating country and year sample collected	3 - alphabetic	HuA (for samples from Brazil 2013)
2	Country code	Country of origin code	3 -alphabetic	Brz
3	DataProducer	Data producer	2 - alphabetic	MB
4	HostPat	Human host and up to 2 pathogens	6 - alphabetic (2 for each organism)	HsViXX (for Human infected with *P. vivax*)
5	SpecimenType	Sample / Specimen type	2 - alphabetic	PS (Plasma)
6	FileSpecificInfo	Information that is specific to this particular file	Variable	20-Manaus-Control-Samples-Results
7	LIMSBC	LIMS barcode	7 - alphanumeric	1234567
8	ext	file extension	Variable - alphabetic	bam

Human Sample Experiment Codes

MaHPIC Human Code	Collection Site and Year	Corresponding MaHPIC Exp#(s)
HuA	Brazil 2013	E09, E12
HuB	Thailand 2015	E08, E09, E10, E12
HuC	Colombia 2015	E09
HuD	BrazilRdj 2015	E09, E12
HuE	NigeriaLagos 2016	E08, E12

HuF	NigeriaKano 2016	E08, E12
HuG	PNG 2016	E09, E12
HuH	Brazil 2016	E09
HuI	Thailand 2016	E09, E12
HuJ	Malaysia 2016	E11, E12

Country Codes

Brazil	Brz
Peru	Per
Malaysia	Mal
Papua New Guinea	Pap
Thailand	Tha
Ghana	Gha
Colombia	Col

Example of file name, for a file derived from a human sample that originated in the city of Manaus, from Brazil:

HuABrzMEHsViXXPS_20-Manaus-Control-Samples-Results_MULTIPL.xlsx

Granularity at which file names are changed VS folder structures are retained
It is to be noted that **NOT ALL** files generated during the course of the MaHPIC project will be renamed. When a clear set of files and / or folder structure is identified, where it would not make sense to extract an individual file from that set or folder, that set of files and / or folders will be packaged as a single entity (zipped file) and made available for download as a whole

Example:
FastQC is a tool that will be run on sequence read data. It produces quality metrics in the form of a text file as well as an html page. The html page depends on content that is present in two folders, named as "Icons" and "Images". We will zip the entire output from the FastQC tool and name the zipped file according to the MaHPIC file naming convention, without touching the internal folders and files. The name of the zip file will help a user in identifying which read set this fastqc file belongs to.

OTHER Notes:

*** Strain confirmation data ***
The time point field will contain "StC" to indicate that this file belongs to Strain Confirmation Data.

Example file name:
E04StCFGMmCyXXIR_endToEnd_XX99-03.bam

*** Reference genome data ***
The Experiment ID field will contain "RfG" to indicate that this file belongs to Reference Genome.

The time point filed will contain numbers indicating the strain or version of the reference genome. It may contain hyphens.
If the reference files belong to “hosts”, the species ID will occupy the first two characters of the <HostPar> field. Instead, if the reference files belong to “pathogens / parasites”, the species ID will occupy the third and fourth characters of the <HostPar> field.
Example file names:
RfG1-0INMmXXXXXX_ORFAll50AA_XXXXXXX.fa.gz
RfG1-0INXXFaXXXX_AnnotatedProteins_XXXXXXX.fa.zip

*** How do we handle multiple strains of same species within 2 letter codes? ***
The file specific info field should be used when there are multiple strains of same species and that info is necessary in the file name.

Core specific notes:
*** Immune Profiling Core ***

- The <TimePoint> field may begin with either ‘T’ or ‘D’. This is to accommodate Innate Immune Profiling core that will receive samples that are not part of the regular time points. They are being called “Daily Time Points” or “Dailies”.
- "Date specimen extracted", when needed by a core such as Innate Immune, will be included in the <FileSpecificInfo> field.
- Based on the above, an example file name, would be as follows:
E13D06IIMmDPXXWB_082613RTi13PANELZ_1234567.fcs